حسین بهشتی فرد / ۹۶۳۱۰۱۱

سوالات تشریحی

سوال ۱)

۱. Unsupervised Learning

یک روش یادگیری ماشین است که در آن الگوریتم بر روی دیتا های بدون لیبل کار می‌کند و الگو ها را شناسایی و یادگیری را انجام می‌دهد. به عنوان یک مثال بخش news گوگل که اخبار های مشابه را دسته بندی می‌کند بر اساس شباهت هایی که دارند.

۲. Supervised Learning

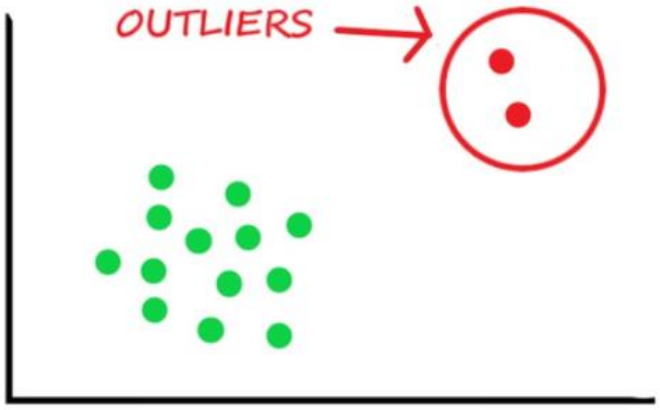
گونه دیگری از روش یادگیری ماشین است که در آن دیتا ها دارای لیبل می‌باشند. در نتیجه با مقایسه با لیبل ها می‌توان یادگیری را انجام داد. این الگوریتم ها برای مسائل رگرسیون و کلاس بندی خوب است. مثلا تشخیص قیمت یک خانه بر اساس متراژ و یا تشخیص اسپم بودن یک ایمیل با توجه به اطلاعات موجود.

۳. Semi-supervised Learning

یک روش دیگر یادگیری ماشین است که در آن مقداری از داده های لیبل دار با داده های بدون لیبل ترکیب می‌شوند. در واقع به این هدف که با استفاده از دیتا های لیبل دار به ماشین یک راهنمایی شود که بتواند به طور دقیق تری یادگیری را انجام دهد.

۴. Outlier

دیتا هایی هستند که در یک نمونه تصادفی از بقیه دیتا ها بسیار دورتر و متفاوت اند. این دیتا ها معمولا نکات مختلفی را نشان می‌دهند. مثلا در مسائلی می‌تواند یک گونه اختلاص و کلاهبرداری را نشان دهد و یا نشان دهنده یک فرد استثنایی.



۵. Dimension

در data Warehousing، dimension مجموعه ای از اطلاعات مرجع در مورد یک رویداد قابل اندازه گیری است. Dimension ها fact های data warehouse را توصیف و دسته بندی می‌کنند که از پاسخ های معنی دار به سوالات تجاری پشتیبانی می‌کند. در واقع dimension یک مجموعه ای از ویژگی های داده است که مورد علاقه تجارت است. ابعاد مواردی مانند مشتری، محصولات، فروشگاه و زمان می‌باشند. برای کاربران data warehouse، ابعاد داده ها نقاطی برای ورود به حقایق عددی هستند (به عنوان مثال فروش ، سود ، درآمد) که یک تجارت می خواهد آنها را مانیتور کند. به معنای عامیانه معمولا به ستون های جدول های دیتابیس(attributes) اشاره دارد. ابعاد و کاهش بعد جز مهم ترین چالش ها هستند.

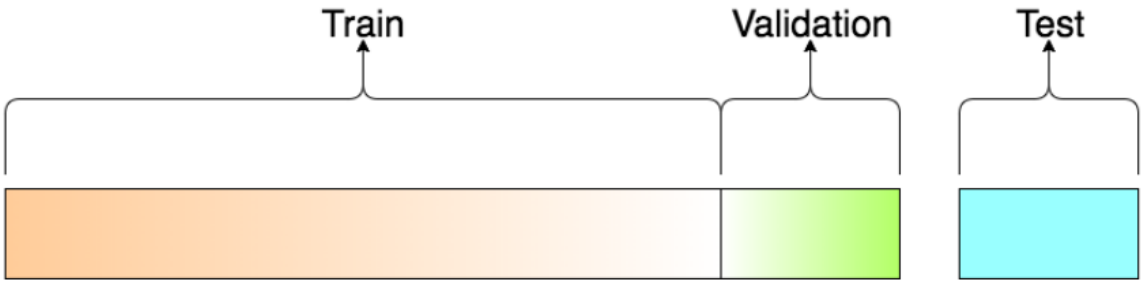
۶. Training, Validating and Testing Data

Training Dataset: مجموعه ای از دیتاهایی است که برای فیت کردن مدل مورد استفاده قرار می‌گیند. یادگیری بر روی آن ها انجام می‌شود و مدل با توجه به این نمونه ها پیش می‌رود و یادگیری را انجام می‌دهد.

Validating Dataset : مجموعه ای از دیتا هایی که برای ارزیابی مدل استفاده می‌شود. از این دیتا های برای یادگیری استفاده نمی‌شود و از آن ها برای تنتظیم hyperparameter های مدل مانند تعداد لایه های مخفی در شبکه های عصبی، استفاده می‌شود. از این دیتا ها برای regularization و early stopping می‌توان استفاده کرد.

Testing Dataset: مجموعه ای هستند که در آخر مدل نهایی که با استفاده از داده های آموزشی تهیه شده است بر روی این دیتا ها به منظور ارزیابی و دقت سنجی مدل انجام می‌شود.

معمولا از ۷۰ و یا ۸۰ درصد داده ها برای آموزش و ۱۰ یا ۱۵ درصد داده برای هر یک از ارزیابی و تست اختصاص داده می‌شود.



A visualization of the splits

۷. Data warehousing

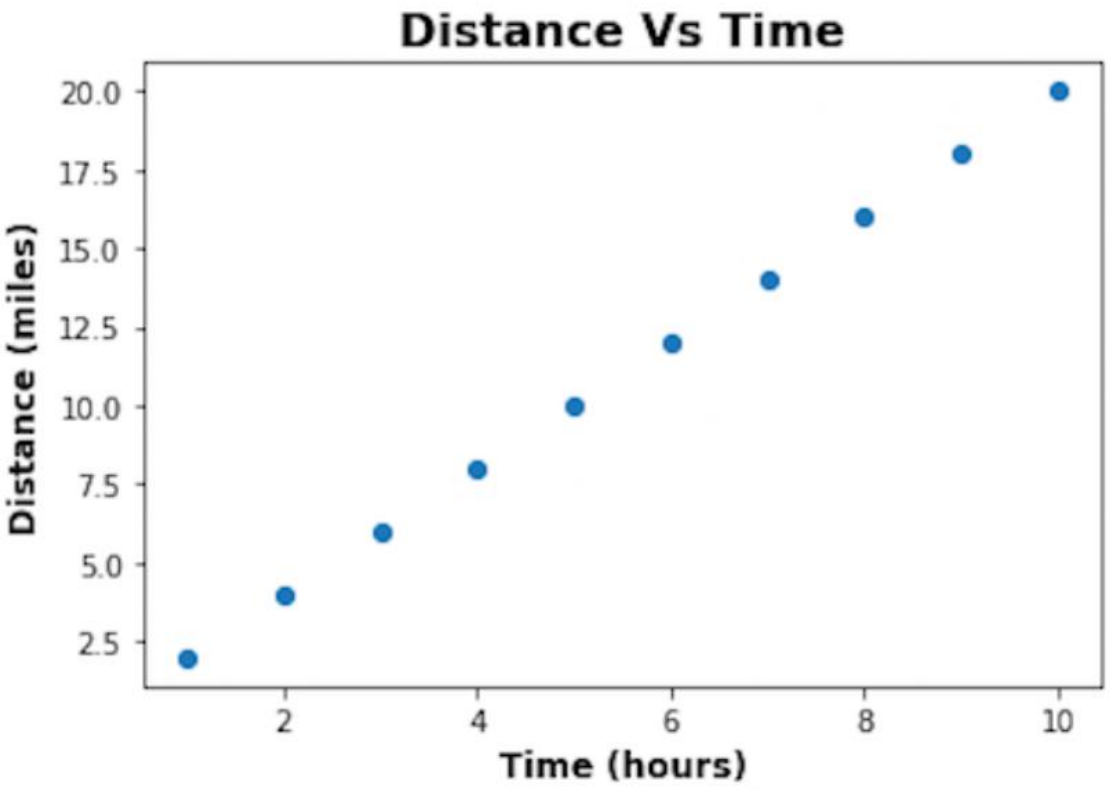
یک ذخیره سازی الکترونیکی از مجموعه عظیمی از داده ها توسط یک شرکت و یا سازمان است. Data warehousing یک بخش حیاتی در هوش تجاری است که از تکنیک های تحلیلی روی داده های تجاری استفاده می‌کنند. در واقع data warehousing فرایند ساخت و استفاده از انبار داده است. یک انبار داده از ادغام داده ها با منبع های متفاوت ناهمگن است.

۸. Missing values

از دست دادن داده ها یک اتفاق معمول است. که این امر عوامل مختلفی را نشان می‌دهد و می‌تواند دلایل متفاوتی داشته باشد. مثلا شاید دیتا در دسترس نبوده است و یا قابل استفاده نبوده یا آن واقعه هنوز رخ نداده است. ممکن است فردی که اطلاعات را پر می‌کرده مقدار مورد نظر را نمی‌دانسته و یا فراموش کرده و دلایل دیگر. در داده کاوی این داده های از دست رفته باید هندل شوند. مرسوم ترین روش ها حذف رکورد های خالی و یا پر کردن آن ها با میانگین گیری از داده های دیگر است.

۹. Independent Variable

متغیر های مستقل متغیر هایی هستند که به متغیر های دیگر وابسته نیستند و معمولا ورودی های فرایندی هستند که در حال تجزیه و تحلیل است. مثلا در یک مسئله اندازه گیری مسافت بر اساس زمان، زمان متغیر مستقل و مسافت وابسته است.



مثالی دیگر که در آن میانگین نمرات یک متغیر وابسته به باقی پارامتر ها است.

year	semester	professor	course	course_title	average_grade	dynamic_learning
2018	Spring	Smith	CS 100	Principles of C++	73	No
2018	Spring	Brown	CS 100	Principles of C++	80	No
2018	Spring	Brown	CS 100	Principles of C++	79	No
2018	Spring	Davis	CS 100	Principles of C++	79	No
2018	Spring	Davis	CS 100	Principles of C++	79	No
2018	Spring	Williams	CS 100	Principles of C++	75	No
2018	Spring	Johnson	CS 100	Principles of C++	76	No
2018	Spring	Jones	CS 100	Principles of C++	76	No
Independent Variables					Dependent Variable	

سوال ۲)

۱. Missing Values Ratio

ستون های دیتا (attributes) با تعداد دیتا های از دست رفته زیاد، یعنی اگر تعداد زیادی از سطر های آن ها خالی باشد، می‌تواند نشان دهنده‌ی اهمیت پایین و زیاد مفید نبودن وجود این ویژگی ها باشد. بنابراین در این روش بر اساس یک حد آستانه ای از دیتا های از دست رفته، بطور کل آن ویژگی حذف می‌شود.

۲. Low Variance Filter

اگر در دیتا های یک ویژگی تغییرات بسیار اندک باشد، نشان می‌دهد که این ویژگی اطلاعات زیادی را شامل نمی‌شود و زیاد مفید نیست(البته بستگی به مسئله دارد). بنابراین در این روش، اگر مقدار واریانس یک ویژگی از یک حد آستانه ای کمتر باشد، آن ویژگی حذف می‌شود. البته در این روش نیاز است که قبل از اعمال، نورمال سازی انجام داد.

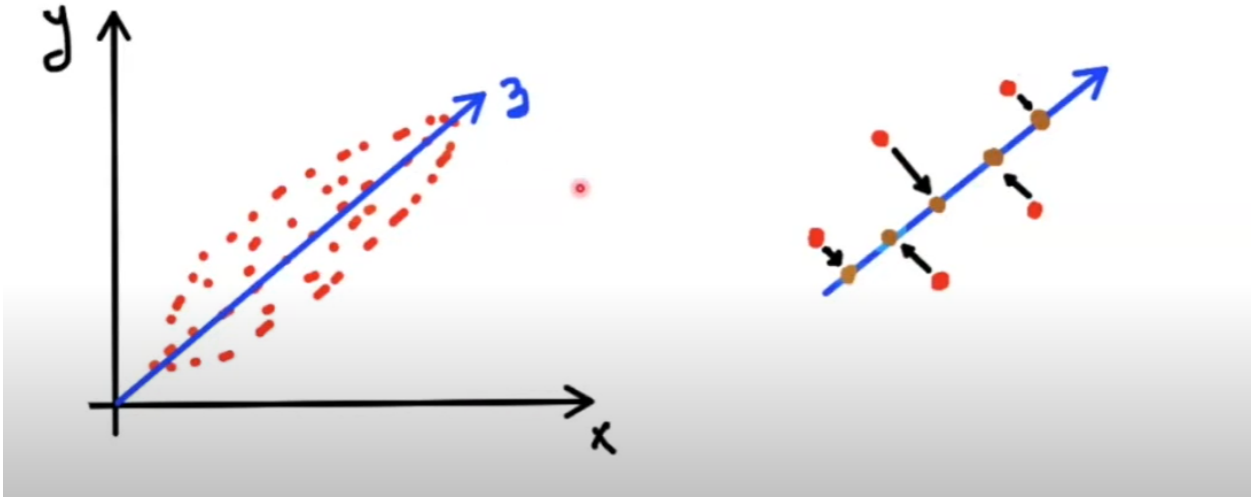
۳. High Correlation Filter

همینطور اگر روند دیتا های دو ویژگی بسیار شبیه باشد، می‌تواند این معنی را بدهد که هر دو ویژگی اطلاعات مشابهی را شامل می‌شوند و وجود یکی از آن ها برای مدل ما کافی است. بنابراین در این روش یکی از این ویژگی ها حذف می‌شود. برای تشخیص این همبستگی می‌توان ضریب همبستگی بین ستون های numerical و nominal را به ترتیب به عنوان Pearson’s Product Moment Coefficient و the Pearson’s chi square محاسبه کنیم. در نهایت جفت ستونی که ضریب همبستگی بالاتر از یک حد آستانه را دارند، حذف کنیم. در این روش نیاز است که قبل از اعمال نرمال سازی انجام شود تا همبستگی معنی دار باشد.

۴. Principal Component Analysis (PCA)

این روش یکی از معروف ترین روش های کاهش بعد است. این روش را با توجه به کورس هایی که دیدم با ذکر یک مثال توضیح می‌دهم.

در این روش هدف تغییر مختصات است. یعنی بجای اینکه یک بعد را تماما حذف کنیم، ابعادی را که همبستگی زیادی دارند با انتقال، بعد را کاهش می‌دهیم. مثلا دیتا های زیر را نگاه کنید. الان اگر یکی از ابعاد x یا y را بخواهیم حذف کنیم تعداد زیادی از دیتا ها از دست می‌روند. در این روش به دنبال حالت شکل راست هستیم که بعد آن کاهش یافته و حدود دیتا ها را دربر دارد. یعنی می‌خواهیم مینیموم مربع خطا را داشته باشیم.

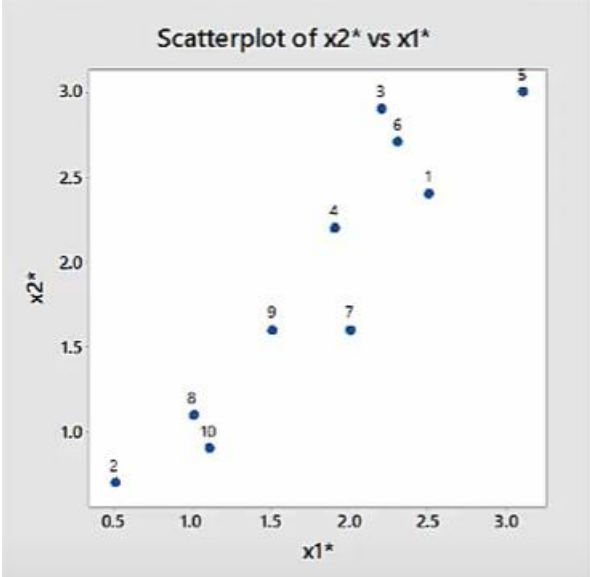


دو ویژگی *x1 و *x2 با دیتا های زیر را در نظر بگیرید

x_1^*	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
x_2^*	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

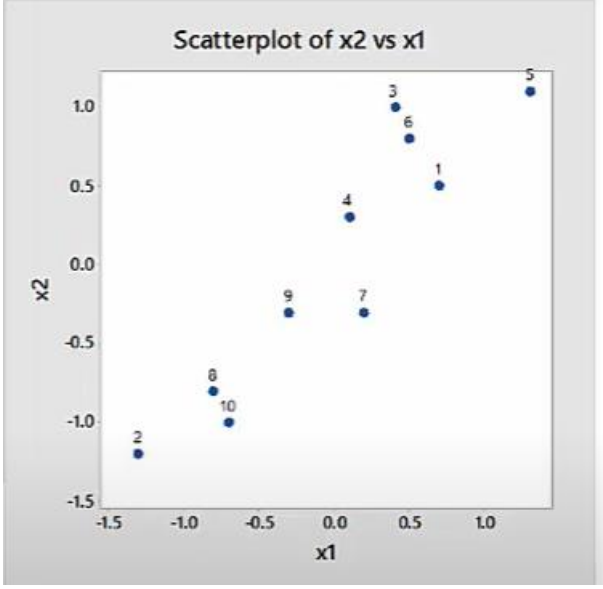
N = 10
p = 2

که نمودار آن بصورت زیر است.



در مرحله اول مرکز مختصات را به (0,0) انتقال می‌دهیم. این عمل را با گرفتن میانگین هر ویژگی و تفریق هر دیتا از جدول با میانگین‌اش انجام می‌دهیم. بعد از این مرحله دیتا ها بصورت زیر خواهند بود.

Recall Step 1: Re-centre data set to origin.											
	x_1	0.69	-1.31	0.39	0.09	1.29	0.49	0.19	-0.81	-0.31	-0.71
	x_2	0.49	-1.21	0.99	0.29	1.09	0.79	-0.31	-0.81	-0.31	-1.01



در مرحله دوم می‌خواهیم که ماتریکس همبستگی را پیدا کنیم که با توجه به اینکه مرکز مختصات (0,0) می‌باشد، فرمول آن بصورت زیر است.

Step 2

Compute the sample variance-covariance matrix **C**.

$$\mathbf{C} = \frac{1}{N-1}(\mathbf{X} - \mathbf{1}\bar{\mathbf{X}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{X}}') = \frac{1}{N-1}\mathbf{X}'\mathbf{X}$$

If the data is standardized, then **C** is the correlation matrix: $\mathbf{C} = \frac{1}{N-1}\mathbf{Z}'\mathbf{Z}$

با وارد کردن دیتا ها و محاسبه داریم:

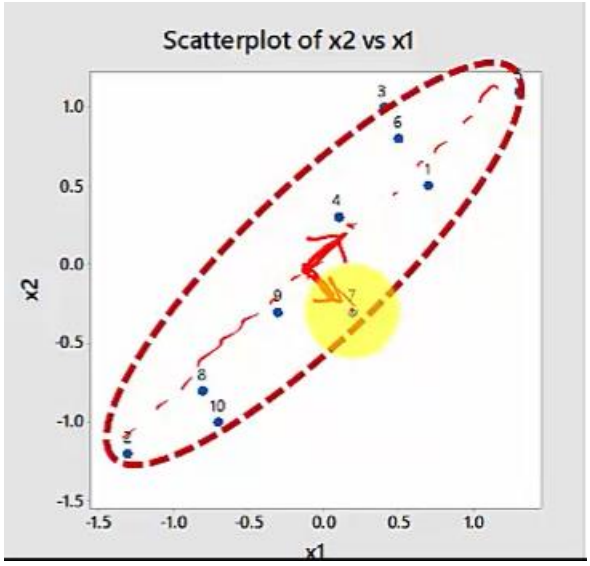
$$C = \frac{1}{10-1}\underbrace{\begin{pmatrix} 0.69 & -1.31 & \cdots \\ 0.49 & -1.21 & \cdots \end{pmatrix}}_{2\times 10}\underbrace{\begin{pmatrix} 0.69 & 0.49 \\ -1.31 & -1.21 \\ \vdots & \vdots \\ \cdot & \cdot \end{pmatrix}}_{10\times 2}$$

$$C = \begin{pmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{pmatrix}$$

Step 3 Compute the eigenvalues λ_i, and (unit or normalised) eigenvectors e_i of **C**, order the corresponding pairs from the highest to the lowest eigenvalues.

در مرحله سوم eigenvalue های هر دو ویژگی را بدست می‌آوریم. بطور کلی eigenvector ها میزان گسترش داده ها روی خط را نشان می‌دهند، مانند شکل زیر:

Variable	Eigenvector 1	Eigenvector 2
x_1	0.678	0.735
x_2	0.735	−0.678
Eigenvalues	1.2840	0.0490
% of total variance	**96.3%**	**3.7%**



Step 4 Choose the components and form the eigenvector matrix **V**.
By ordering the eigenvectors according to the eigenvalues, this gives the components in order of their significance. Hence, the eigenvector with the highest eigenvalue is the principal component. The components of lesser significance can be ignored, so as to reduce the dimensions of the data set.

در مرحله چهارم pc1 را حفظ می‌کنیم زیرا بیش از ۹۶ درصد واریانس را شامل می‌شود. بنابراین:

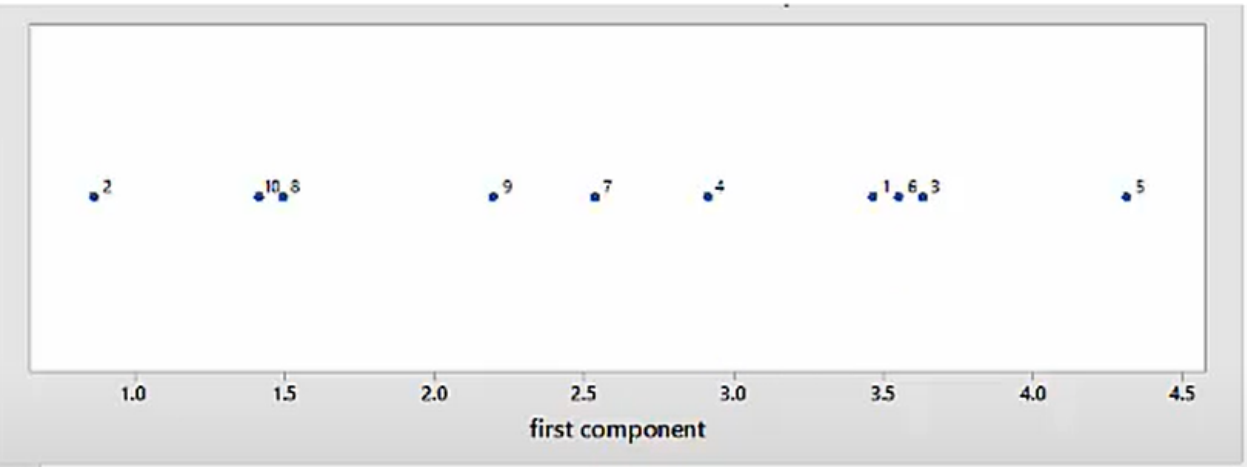
$$\mathbf{V} = \begin{pmatrix} 0.678 \\ 0.735 \end{pmatrix}$$

Step 5 Derive the new data set by taking **Y = XV.**
Basically we have transformed our data so that it is expressed in terms of the patterns between them, where the patterns are the lines that most closely describe the relationships between the data.

در مرحله آخر ماتریس V را در ماتریس دیتا های خود ضرب می‌کنیم.

$$Y = \begin{bmatrix} 2.5 & 2.4 \\ 0.5 & 0.7 \\ 2.2 & 2.9 \\ \vdots & \vdots \\ 1.1 & 0.9 \end{bmatrix} \begin{bmatrix} 0.678 \\ 0.735 \end{bmatrix} \quad \therefore Y = \begin{bmatrix} 3.459 \\ 0.854 \\ 3.623 \\ \vdots \\ 1.407 \end{bmatrix}$$

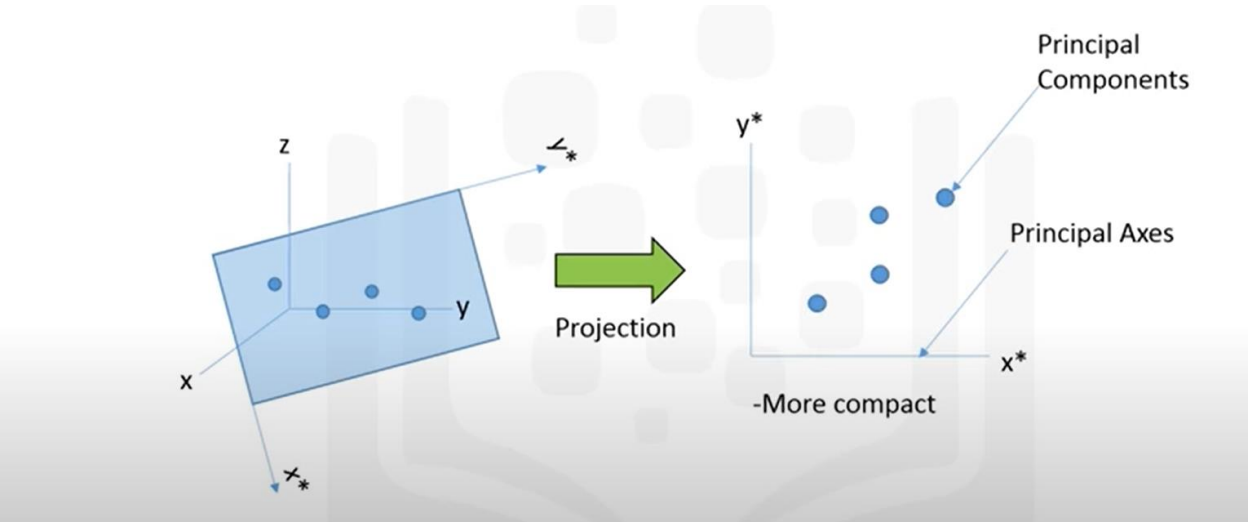
در نتیجه نمودار کاهش بعد پیدا می‌کند و از یک نمودار دو بعد به یک نمودار تک بعدی بصورت زیر تبدیل می‌شود که دیتا ها به این خط انتقال پیدا کرده اند.



تفاوت انتخاب و استخراج ویژگی:

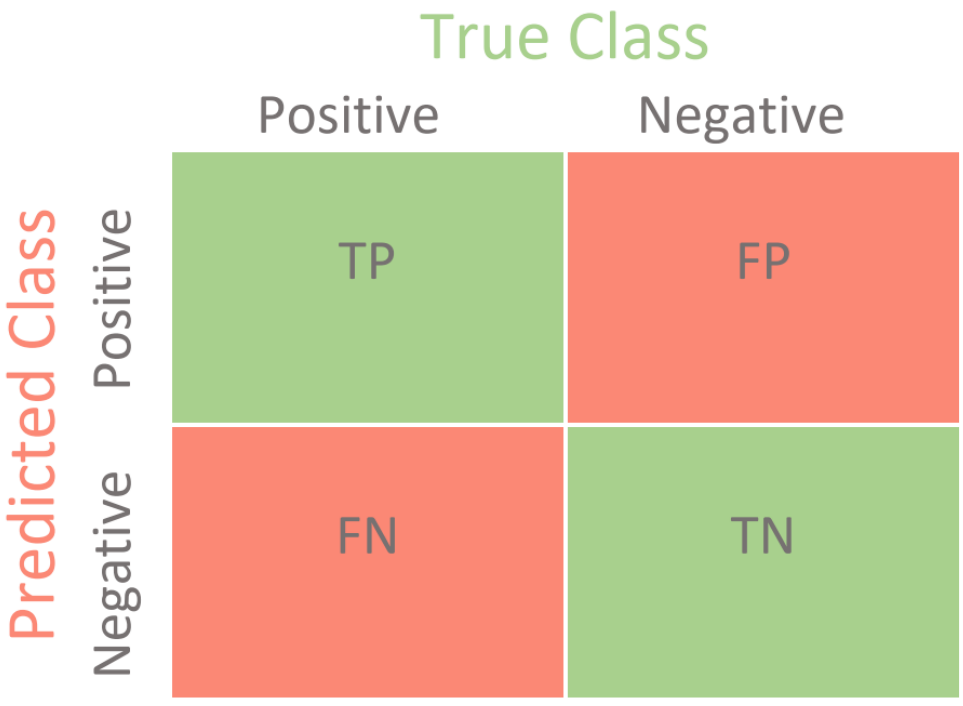
در انتخاب ویژگی ما با استفاده از wrapper ها، filter ها و embedded تعدادی از ویژگی ها را کاملا حذف می‌کنیم. مانند سه روش اولی که گفته شد.

اما در استخراج ویژگی ما از روش هایی مانند انتقال مختصات مانند روش PCA که توضیح داده شد استفاده می‌کنیم و هیچ ویژگی ای را دور نمی‌ریزیم. مانند شکل زیر:



سوال ۳)

ترم ها در ماتریکس بهم ریختگی بصورت زیر هستند:



true positives (TP)

زمانی که درست بوده اند و ما نیز YES پیش بینی کرده ایم.

true negatives (TN)

زمانی که درست نبوده اند و ما نیز NO پیش‌بینی کرده ایم.

false positives (FP)

زمانی که درست نبوده اند و ما YES پیش‌بینی کرده ایم.

false negatives (FN)

زمانی که درست بوده اند و ما NO پیش‌بینی کرده ایم.

Precision

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

همانطور که می‌بینیم نسبت تعداد پیش‌بینی هایی که درست بوده اند و ما مثبت اعلام کرده ایم به تعداد کل پیش‌بینی های مثبتی که انجام دادیم. در واقع میزان دقت در پیش‌بینی هایی که YES اعلام کردیم را نشان می‌دهد. Precision برای زمانی خوب است که هزینه FP زیاد باشد مثلا در تشخیص اسپم بودن یک ایمیل اگر این نسبت کم باشد ممکن است ایمیل های مهمی از دست برود.

Recall

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

فراخوانی نسبت پیش‌بینی هایی که مثبت بوده اند و ما نیز YES اعلام کرده ایم به تعداد کل مواردی که مثبت بوده اند. یعنی در واقع نشان می‌دهد که الگوریتم ما چند درصد مواردی که واقعا مثبت بوده اند را درست پیش‌بینی کرده است. این معیار برای زمانی خوب است که هزینه FN زیاد است. مثلا وقتی که یک شخصی بیمار باشد و ما NO اعلام کنیم هزینه سنگینی دربر خواهد داشت.

F1-Score

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

فرمول امتیاز اف بصورت بالا است و این معیار برای زمانی نیاز است که می‌خواهیم یک بالانس بین دقت و فراخوانی پیدا کنیم.

سوال ۴)

فرمول کواریانس و همبستگی بصورت زیر است.

$$Cov(X,Y) = E[XY] - E[Y]E[X]$$

$$\rho(X,Y) = \frac{Cov(X.Y)}{\sqrt{Var(X)\,Var(Y)}}$$

اگر دو متغیر مستقل باشند با توجه به اینکه E[XY] = E[Y]E[X]، همبستگی بین آن ها صفر می‌شود اما اگر همبستگی بین دو متغیر صفر باشد صرفا دو متغیر وابسته خطی نیستند و نمی‌توان بیان کرد که که مستقل از یکدیگر اند. بنابراین این دو متغیر می‌توانند بصورت غیر خطی با یکدیگر مرتبط باشند که همبستگی نمی‌تواند آن را تشخیص دهد.

به مثال زیر توجه کنید. در این مثال correlation برابر صفر می‌شود اما می‌بینیم که X و Y مستقل از هم نیستند.

$$P(X = x) = 1/3 \text{ for } x = -1, 0, 1 \text{ and } Y = X^2.$$

سوال ۵)

Data Cleaning

این مرحله بطور کلی شامل پر کردن دیتا های از دست رفته، هندل کردن دیتا های نویز دار، شناسایی یا حذف outlier ها و برطرف کردن ناسازگاری ها می‌باشد.

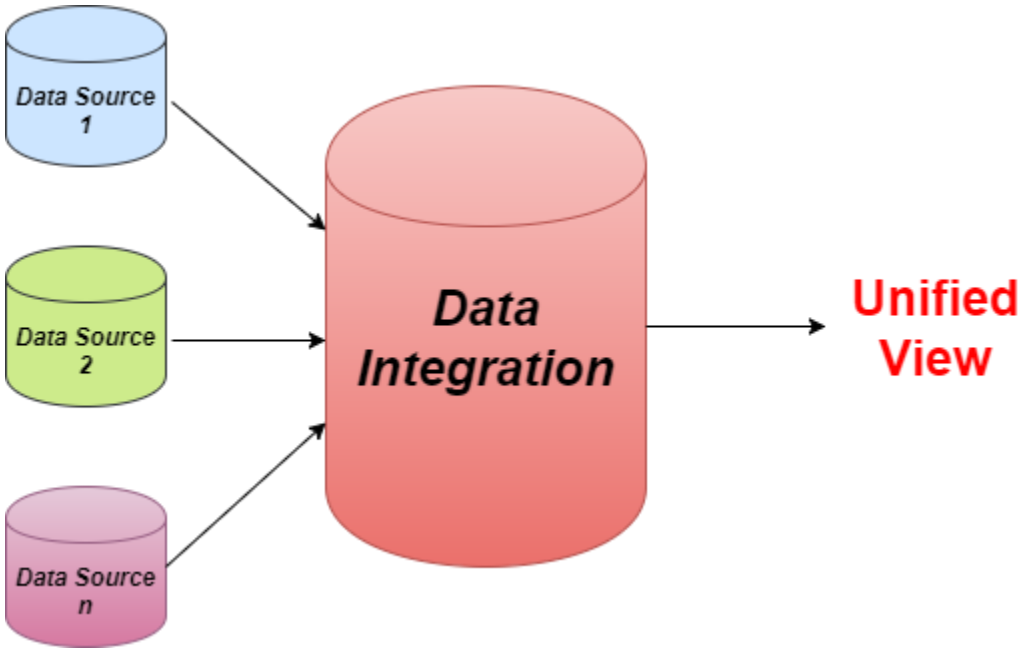
همانطور که قبلا اشاره کردیم، از دست رفتن و نداشتن دیتا ها می‌تواند دلایل متعددی داشته باشد. ممکن از دیتا ها در آن زمان در دسترس نبوده است، آن واقعه رخ نداده است مانند فیلد میزان درامد برای کسی که شغلی ندارد، ممکن است با سایر رکورد ها تداخل داشته و حذف شده و یا طی اتفاقات دیگری وارد نشده و یا حذف شده اند. برای پردازش روی داده ها، این رکورد های خالی می‌توانند مشکل ایجاد کنند برای همین معمولا در پیش پردازش داده ها این فیلد ها با مقادیری مانند یک مقدار یکسان برای همه که نشان می‌دهد داده در دستریس

نیست یا گرفتن میانگین دیگر رکورد ها و وارد کردن آن و یا اگر تعداد ویژگی ها زیاد است و تعداد زیادی از رکورد های مربوط به یک attribute خالی است، ممکن است بطور کلی آن فیلد حذف گردد.

همینطور مشکلات دیگری مانند نویز داشتن و صحیح نبودن دیتا ها می‌تواند وجود داشته باشد که می‌تواند عوامل متعددی مانند محدودیت های تکنولوژی،جابه‌جایی داده ها، ناسازگاری و ... داشاه باشد که در این مرحله باید هندل شوند تا بتوانیم نتیجه خوبی بگیریم. معمولا برای هندل کردن این مشکلات از تکنیک هایی مانند Binning، regression و clustering استفاده می‌کنند.

Data Integration

ادغام دیتا ها یک فرایند است که در آن داده ها از منابع ناهمگون متفاوت که هر کدام ممکن است چندین دیتابیس و پرونده داشته باشد را ترکیب می‌کنند بصورتی که انسجام داشته باشد و یک شکل یکسان داشته باشند.



بطور کلی دو رویکرد برای ادغام دیتا ها وجود دارد.

Tight Coupling: در این رویکرد دیتا ها از منابع مختلف از طریق فرایند ETL(استخراج، بارگیری و لودینگ) در یک مکان فیزیکی ترکیب می‌شوند و این انبار داده بصورت یک کامپوننت بازیابی اطلاعات کار می‌کند.

Loose Coupling: در این روش دیتا ها در محل خود باقی می‌مانند. در عوض یک اینترفیس ساخته می‌شود که کوئری را از کاربر می‌گیرد و آن را بطوری که قابل فهم باشد برای دیتابیس تغییر می‌دهد و کوئری را به منبع مربوطه می‌فرستد و دیتا ها را برمی‌گرداند.

فرایند ادغام دیتا ها مشکلات و چالش های خود را دارد. از جمله این چالش ها متفاوت بودن schema ها و اختلافات و ناسازگاری دیتا ها، تکراری بودن دیتا ها و redundancy داشتن می‌باشد.

Data Transformation

هدف از این فرایند تبدیل دیتا ها به یک فرم مناسب و آماده سازی آن ها برای عملیات داده کاوی می‌باشد. ۴ عملیات زیر از جمله عملیات های مهم این فرایند می‌باشند.

Normalization: نیاز است که داده ها نورمال سازی بشوند تا یک فرم و مقیاس یکسان داشته باشند.

Attribute selection: در این عملیات از روی ویژگی های موجود، ویژگی های جدیدی که به کار داده کاوی کمک می‌کند، ساخته می‌شود.

Discretization: این مرحله به منظور جایگزینی داده های عددی خام با سطوح بازه ای و مفهومی می‌باشد.

Concept Hierarchy Generation: در این عملیات ویژگی ها به سطوح بالا تر تبدیل می‌شوند مثلا به عنوان مثال تبدیل ویژگی شهر به کشور.

سوال ۶)

تراکنش ها بصورت زیر داده شده است.

ID	Items
T1	نان، الویه، پنیر
T2	نان، الویه
T3	نان، کره، مربا
T4	مربا، کره
T5	مربا، پنیر
T6	نان، کره، مربا

در ابتدا مجموعه های تکی را محاسبه می‌کنیم:

Item	Sup %

نان	$\frac{4}{6} * 100 \cong 66/66٪$
الویه	$\frac{2}{6} * 100 \cong 33/33٪$
پنیر	$\frac{2}{6} * 100 \cong 33/33٪$
کره	$\frac{3}{6} * 100 = 50٪$
مربا	$\frac{4}{6} * 100 \cong 66/66٪$

همینطور که محاسبه کردیم مقدار support تمامی آیتم ها از حد آستانه پشتیبان داده شده بیشتر است، بنابراین هیچ یک از آیتم ها را حذف نمی‌کنیم.

حال مجموعه های دوتایی را محاسبه می‌کنیم:

ItemSet	Sup %
نان، الویه	$\frac{2}{6} * 100 \cong 33/33٪$
نان، پنیر	$\frac{1}{6} * 100 \cong 16/66٪$
نان، کره	$\frac{2}{6} * 100 \cong 33/33٪$
نان، مربا	$\frac{2}{6} * 100 \cong 33/33٪$
الویه، پنیر	$\frac{1}{6} * 100 \cong 16/66٪$
الویه، کره	$\frac{0}{6} * 100 = 0$
الویه، مربا	$\frac{0}{6} * 100 = 0$
پنیر، کره	$\frac{0}{6} * 100 = 0$
پنیر، مربا	$\frac{1}{6} * 100 \cong 16/66٪$

کره، مربا	$\frac{3}{6} * 100 = 50\%$

مجموعه هایی که مقدار support آن ها از آستانه کمتر است را حذف می‌کنیم و از روی مجموعه های باقی مانده، مجموعه آیتم ها ۳تایی را می‌سازیم. البته در اینجا برای اینکه جدول طولانی نشود با توجه به حالات دوتایی آن حالاتی را که می‌دانیم احتمالشان کمتر از آستانه شد، آن ترکیب را دیگر درنظر نمی‌گیریم چون در مجموعه های سه تایی این احتمال کمتر مساوی است. در واقع داریم حرص انجام می‌دهیم.

ItemSet	Sup %
نان، کره، مربا	$\frac{2}{6} * 100 \cong 33/33\%$

طبق محاسبات انجام شده با توجه به آستانه پشتیبانی تعیین شده، تنها مجموعه {نان، کره، مربا} با مقدار پشتیبانی $33/33\%$ یک مجموعه پرتکرار است.

حال قواعد انجمنی را نوشته و میزان اطمینان را نشان می‌دهیم.

فرمول محاسبه confidence بصورت زیر می‌باشد.

$$Rule:\ X \Rightarrow Y \longrightarrow Confidence = \frac{frq(X,Y)}{frq(X)}$$

Rules	Confidence %
{نان ،کره} → {مربا}	$\frac{2}{4} * 100 = 50\%$
{نان ،مربا} → {کره}	$\frac{2}{3} * 100 \cong 66/66\%$
{مربا ،کره} → {نان}	$\frac{2}{4} * 100 = 50\%$
{نان} → {مربا،کره}	$\frac{2}{3} * 100 \cong 66/66\%$
{مربا} → {نان، کره}	$\frac{2}{2} * 100 = 100\%$
{کره} → {مربا ،نان}	$\frac{2}{2} * 100 = 100\%$

با محاسبات انجام شده می‌بینیم که طبق آستانه اطمینان تعیین شده، دو قانون انجمنی {مربا} → {نان، کره} و {کره} → {مربا ،نان} با درصد اطمینان ۱۰۰ درصد و دو قانون {نان ،مربا} → {کره} و {نان} → {مربا،کره} با درصد اطمینان ۶۶/۶۶ درصد باقی می‌مانند.

سوال ۷)

الف)

در ابتدا مجموعه های تکی را محاسبه می‌کنیم:

Item	Sup %
نان	$\frac{4}{6} * 100 \cong 66/66٪$
الویه	$\frac{2}{6} * 100 \cong 33/33٪$
پنیر	$\frac{2}{6} * 100 \cong 33/33٪$
کره	$\frac{3}{6} * 100 = 50٪$
مربا	$\frac{4}{6} * 100 \cong 66/66٪$

همینطور که محاسبه کردیم مقدار support تمامی آیتم ها از حد آستانه پشتیبان داده شده بیشتر است، بنابراین هیچ یک از آیتم ها را حذف نمی‌کنیم.

حال اقلام را بر اساس بیشترین تکرار مرتب می‌کنیم:

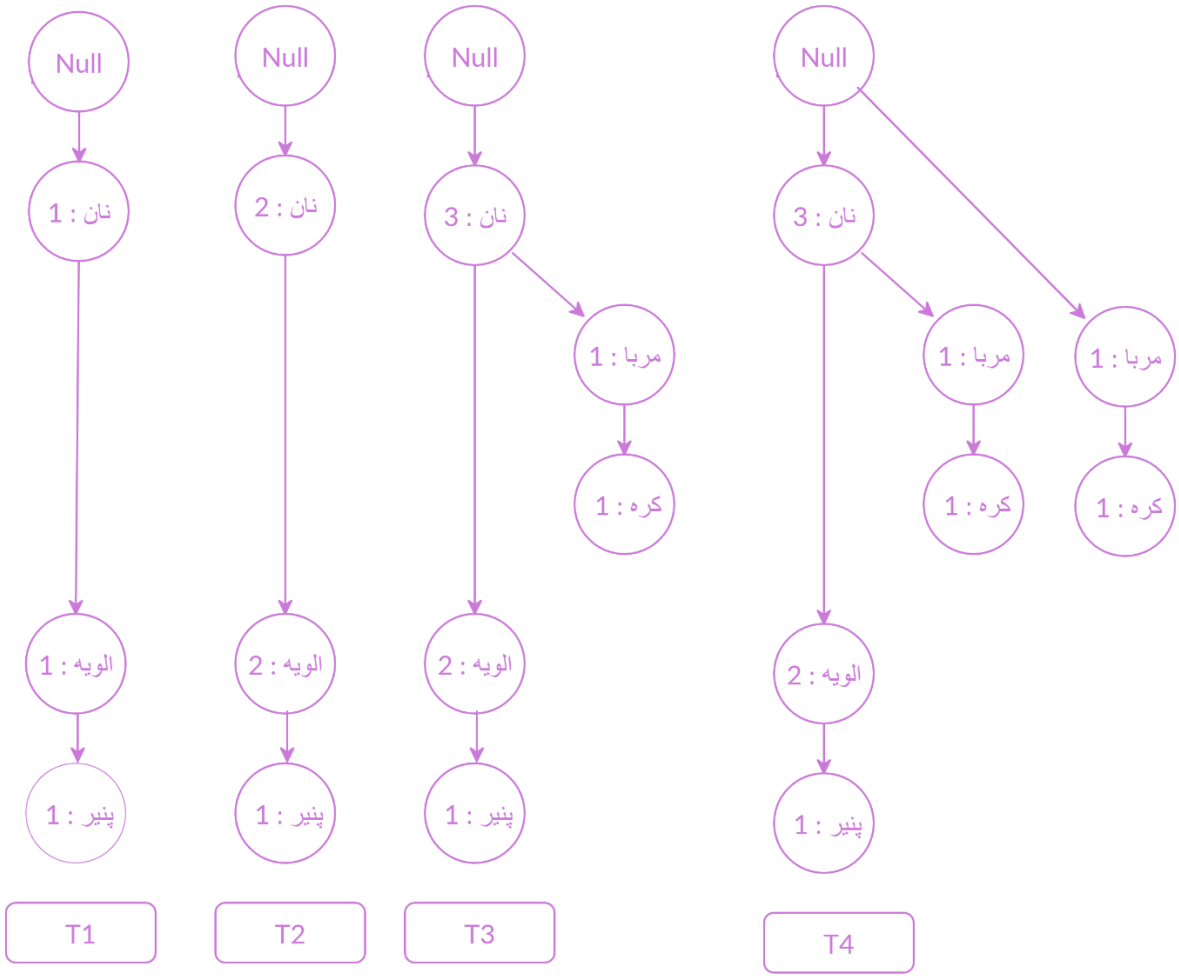
Item	freq
نان	4
مربا	4
کره	3
الویه	2

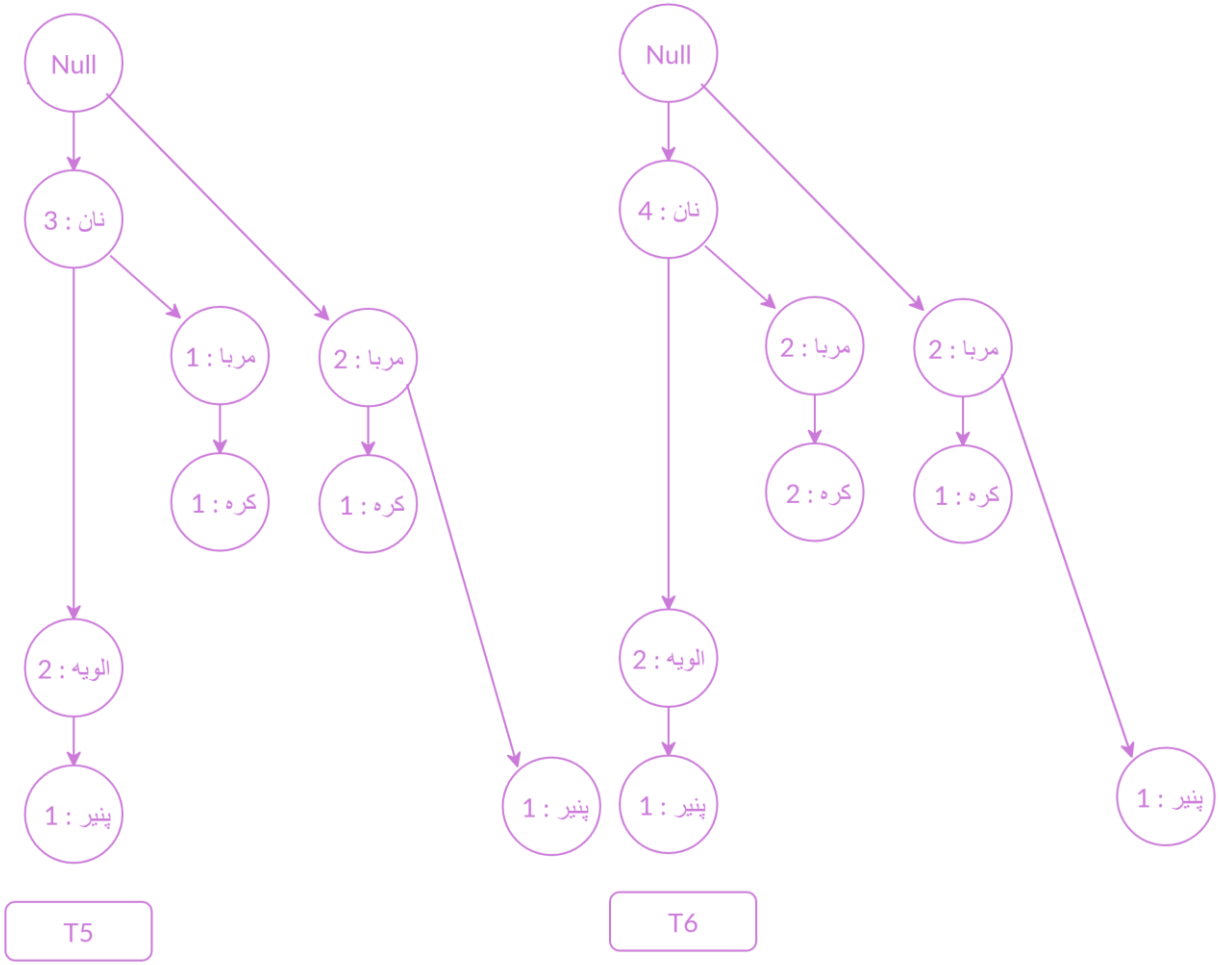
2	پنیر

در مرحله بعد مجموعه ها را بر اساس میزان پشتیبانی اقلام مرتب می‌کنیم.

ID	Items	Ordered Items
T1	نان، الویه، پنیر	پنیر، الویه، نان
T2	نان، الویه	الویه، نان
T3	نان، کره، مربا	کره، مربا، نان
T4	مربا، کره	کره، مربا
T5	مربا، پنیر	پنیر، مربا
T6	نان، کره، مربا	کره، مربا، نان

حال بر اساس تراکنش ها درخت الگو پر تکرار را می‌سازیم.


Null
1 : نان
1 : الویه
1 : پنیر
T1
Null
2 : نان
2 : الویه
1 : پنیر
T2
Null
3 : نان
1 : مربا
1 : کره
2 : الویه
1 : پنیر
T3
Null
3 : نان
1 : مربا
1 : کره
1 : مربا
1 : کره
2 : الویه
1 : پنیر
T4



ب)

Items(Descending)	Conditional Pattern Base	Conditional FP tree
پنیر	{نان، الویه : 1}، {مربا : 1}	-
الویه	{نان : 2}	{نان : 2}
کره	{نان، مربا : 2} ، {مربا : 1}	{مربا : 3}
مربا	{نان : 2}	{نان : 2}
نان	-	-

Frequent patterns: < نان، الویه : ۲ > , < مربا، کره : ۳ > , < نان، مربا : ۲ >